DOCUMENT RESUME

ED 155 214 TM 007 217

AUTHOR Marsh, Herbert W.; And Others
TITLE The Validity of Students' Evaluations of Instructional Effectiveness: A Comparison of Faculty Self-Evaluations and Evaluations by Their Students.
PUB DATE May 78
NOTE 30p.; Paper presented at the Annual Meeting of the Association for Institutional Research (Houston, Texas, May, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *College Teachers; Correlation; Course Evaluation; Effective Teaching; Evaluation Criteria; Higher Education; *Self Evaluation; Social Sciences; *Student Evaluation of Teacher Performance; Student Teacher Relationship; *Teacher Evaluation; *Teaching Quality; Test Reliability; *Test Validity; Undergraduate Students

ABSTRACT

Student evaluations of teacher effectiveness have been accepted by instructors as helpful indicators of performance, but their validity and use in tenure and promotion decisions has been questioned by faculty. Students and instructors in 207 social science courses completed evaluations of instructional effectiveness at the conclusion of the semester. Each of the 65 participating faculty members designated the course in which his or her teaching had been the most and the least effective. The instuctors then evaluated their teaching in both courses. Instructor and student evaluations contained identical items, samples of which are appended. Faculty and students agreed upon six factors of teacher effectiveness: breadth of coverage, organization, group interaction, individual interaction, instructor enthusiasm and learning/value. Factor analysis revealed that student and faculty agreement on evaluation factors was high. Student evaluation of the courses designiated most effective by instructors was higher on all scores. The median evaluation was the same for both groups. The study indicated that self-evaluation is beneficial to faculty, and that student evaluation of teaching effectiveness is a valid process worthy of faculty confidence. (Author/JAG)

The Validity of Students' Evaluations of Instructional Effectiveness: A Comparison of Faculty Self-Evaluations and Evaluations by Their Students

Herbert W. Marsh
University of Southern California

Jesse U. Overall
California State College, Dominguez Hills

Steven P. Kesler
University of Southern California

This paper was the basis of a talk presented at the
Annual Meeting of the ASSOCIATION FOR INSTITUTIONAL RESEARCH
Houston, May, 1978.

Running Head: Faculty Self-Evaluations

Abstract

Faculty who taught two courses during the spring 1976 semester evaluated their own teaching in each of the two courses as well as being evaluated by their students. These faculty felt that quality of teaching should be given more importance and that students' evaluations were useful for the faculty themselves, but expressed reservations about both the validity of the student ratings and their use in tenure/promotion decisions. In spite of these reservations, there was considerable student-faculty agreement. Separate factor analyses of the two sets of ratings both supported the six evaluation factors which had previously been identified, indicating student-faculty agreement on the dimensions which underlie ratings of effective teaching. Validity coefficients, correlations between student and faculty ratings on the same evaluation factors varied between .33 and .67 (median r = .49). The difference between mean faculty self-evaluations, averaged across faculty, and mean students' evaluations were small; the median evaluation was the same for both groups. Students and faculty agreed upon the teaching behaviors which were more descriptive and less descriptive of the faculty as a whole (r = .77). Finally, when faculty indicated that their teaching was "most effective" in one course and "less effective" in another, students' evaluation of the "most effective" courses were higher on all evaluation scores. These findings offer strong new support for the validity of students' evaluations, suggest the possible usefulness of faculty self-evaluations, and should be particularly helpful in overcoming faculty resistance to the use of students' evaluations.

The Validity of Students' Evaluations of Instructional Effectiveness: A Comparison of Faculty Self-Evaluations and Evaluations by Their Students

Students' evaluations of instructional effectiveness continue to be both widely used and controversial. Decreasing enrollments and an increased emphasis on accountability, often externally motivated, have brought about a renewed interest in evaluating quality of instruction. Consequently, students' evaluations--often the only measure of teaching effectiveness which are regularly available--not only impact on faculty self-esteem and reputation, but may also affect their careers. However, demonstrations of the validity of the students' evaluations are generally limited to specialized settings or employ criteria which can be easily challenged. Faculty, like other human beings, are suspicious of the processes used to evaluate them. As long as there remains broad faculty distrust of the validity of students' evaluations, their usefulness will be severely limited. The purpose of this study is to show that faculty self-evaluations of their own teaching, in spite of faculty reservations about the validity of the student ratings, show good agreement with students' evaluations of teaching effectiveness.

The most common criticism of students' evaluations, besides the feeling that they lack validity, is that they are biased by variables unrelated to teaching effectiveness. There is considerable evidence that most background variables such as class size, reason for taking the course, workload and grade point average have little relationship to students' evaluations (Marsh 1978; Marsh, Overall & Thomas, 1976; McKeachie, 1973; Remmers, 1963). However, this apparent lack of bias does not necessarily mean that students' evaluations are also valid measures of instructional quality. Validating students' ratings is difficult because there are no clearly defined criteria of instructional quality. Indeed, validating the measurement of any complex

construct like effective teaching requires the use of many alternative criteria.

Validity studies have typically used student performance on standardized examinations as a criterion. When different instructors teach different sections of the same course, the sections which evaluate their instructor most highly are also the sections that perform better on the standardized examinations given to all sections (Overall & Marsh, 1978, Centra, 1977; Marsh, Fleiner & Thomas, 1975; Frey, 1973; Cohen & Berger, 1970; Morsh, Burgess & Smith, 1955), though Rodin and Rodin (1972) did report negative findings. Overall and Marsh (1978) also showed the affective consequences of a course (feelings of course mastery and disposition to pursue the subject further) were significantly related to students' evaluations. Marsh (1977), considering an alternative criterion, asked graduating seniors to nominate "most outstanding" and "least outstanding" faculty in a separate survey which was mailed to them. Classroom evaluations of the "most outstanding" faculty were consistently much more favorable than the evaluations of the "least outstanding" faculty teaching similar classes. This implies that the qualities of an instructor which cause him to be nominated by graduating seniors are also reflected in classroom evaluations.

Obvious criteria for validating students' evaluations are the corresponding evaluations made by the faculty themselves (self-evaluations) or the evaluations made by other faculty (colleague evaluations). Morsh , Burgess and Smith (1955) collected evaluations of instruction in an Air Force Training Program from students, colleagues, and supervisors in addition to measuring actual student learning on a standardized examination. Although there was a strong positive relationship between students' evaluations and actual learning ($r = .49$), there was no significant relationship between actual learning and either colleague or supervisor evaluations. Guthrie (1954), Maslow

and Zimmerman (1956) and Blackburn and Clark (1975) have all reported high to moderate correlations between colleague ratings and student ratings (correlations ranging from .43 to .69 on ratings of overall teacher effectiveness). However, none of the studies actually required colleagues to base their evaluations on visits to the classroom, and it is likely that at least part of the basis of the colleague evaluations was feedback from students.

Centra (1975) compared colleague and student evaluations in a setting which reduced the probable confounding of the two sources of information. Colleague evaluations were based upon actual classroom visitation, and the study was conducted in a university at which teaching reputations had not yet been established. Each of 54 faculty was evaluated twice by each of three different colleagues. While there was good agreement between the evaluations of the same colleague on different visits ($r = .78$), there was little agreement between the evaluations of different colleagues ($r = .26$). The lack of reliability in the colleague ratings precluded any good correspondence with student evaluations and the median correlation between the two groups for 16 evaluation items was only .20.

Blackburn and Clark (1975) reported a correlation of only .19 between faculty self-evaluations and student evaluations. However, the faculty self-evaluations were only general impressions of teaching effectiveness which were not tied to actual performance in a particular course, while the student evaluations were based upon actual teaching. Centra (1972) asked faculty to select a single course in which to evaluate themselves and to be evaluated by their students. Faculty self-evaluations of their selected course tended to be more favorable than the evaluations by their students and showed only

modest correlations with student ratings; the median correlation for the 17 evaluation items was r = .21. This indicates that faculty who saw themselves as most effective were also evaluated somewhat more favorably by their students. However, items which received consistently high or low ratings by all faculty were also given similar ratings by all students. The correlation between faculty mean responses to the evaluation items and student mean responses to the same items was .77. While faculty and students showed only modest agreement on which faculty were most effective teachers, there was good agreement on the behaviors at which faculty as a whole are best and worst.

Previous research has found only modest correspondence between faculty self-evaluations and student evaluations. However, faculty evaluating their own teaching in a general sense, or even in one specific course, are not forced to differentiate between their own more and less effective teaching. Students, on the other hand, have a wide basis of comparison against which to evaluate the performance of any given instructor. In the present study, faculty who had just completed teaching two different courses were asked to indicate in which course their teaching was more effective and in which less and to evaluate their teaching in both courses. This procedure assured that faculty self-evaluations were based upon teaching in a specific course and forced faculty to differentiate between their own more and less effective teaching.

METHOD

Students' Evaluations and Survey Instrument

During the Spring 1976 semester at the University of Southern California, students' evaluations were collected in a total of 207 undergraduate courses which were taught by faculty in the Division of Social Sciences. Graduate level courses and courses taught by Teaching Assistants were not included in these analyses. Student evaluation instruments were sent to faculty in charge of all courses several weeks before the end of the semester and were actually used in virtually all of these courses. The evaluation forms were administered during a class period prior to the final examination, collected by a student in the class, and immediately taken to the department office. An average of 78% of the students enrolled in these courses completed the survey forms.

The evaluation instrument consisted of 24 evaluation items adapted from Hildebrand, Wilson and Dienst (1971) and several additional background/demographic variables. The reliability of individual evaluation items (Marsh, 1976b), based upon sets of responses from 20 students in each class, varied between .73 and .90 (median .84). Coefficient alphas, determined for both students' evaluations and faculty self-evaluations as part of this study, were computed according to Method 2 of the Statistical Package for Social Sciences (Nie et al., 1977).

Both the students' evaluations and the faculty self-evaluations were summarized by eight evaluation scores, factor scores representing the six evaluation factors and overall ratings of the teacher and the course. Evaluation factor scores were weighted averages of standardized items. The weights, factor score coefficients, were derived from a factor analysis (Nie et al., 1975) done across all courses which were evaluated during a two semester period of time (Marsh, 1976a). The evaluation scores are characterized as follows:

BREADTH OF COVERAGE--Presents a broad background encompassing alternative approaches to the subject and emphasizing analytic ability and conceptual understanding.

ORGANIZATION--Is well organized and prepared, giving explanations and answers which are clear.

GROUP INTERACTION--Encourages class discussions and invites students to share their own ideas or be critical of those presented by the instructor.

INDIVIDUAL INTERACTION--Is friendly and interested in students and is accessible to them.

INSTRUCTOR ENTHUSIASM--Displays enthusiasm, energy, humor and ability to hold student interest.

LEARNING/VALUE--The extent to which students experienced a valuable learning experience which was intellectually demanding.

OVERALL INSTRUCTOR--A single item asking "How does this instructor compare with other instructors you have had at this school?"

OVERALL COURSE--A single item asking "How does this course compare with other courses you have had at this school?"

Faculty Self-Evaluations and Survey Instrument

During the 1976 Spring Semester 65 different instructors in the Division of Social Sciences taught at least two courses in which they were also evaluated by their students. A Faculty Self-evaluation survey was sent to these teachers at the end of the term, but before summaries of the students' evaluations had been returned. Faculty were assured that their responses would remain confidential. Instructors indicated in which course their teaching was "most effective" and in which "less effective" and rated the difference in effectiveness. Faculty then evaluated both courses with a set of items which were identical to those used by students except that they were worded in the first person. Faculty were asked to rate their own teaching effectiveness and not to report how students would rate them. The survey form also contained additional items related to their attitudes towards students' evaluations and selected background/demographic variables.

A total of 51 (78%) surveys, including evaluations of 83 different undergraduate courses, were returned in response to the original survey and two additional mailings to non-respondents. Thirty-two of the respondents evaluated two undergraduate courses which they designated to be "most effective" or "less effective". The remaining 19 respondents either evaluated one undergraduate course and one graduate level course or evaluated only one undergraduate level course.

RESULTS

Factor Analysis

Factor analysis is used to describe the underlying dimensions which are actually being measured by a set of questions. The technique identifies clusters of items which are highly related to each other and less related to other clusters of items. The simple structure criterion for factor analysis attempts to determine dimensions so that any given item loads high on one dimension and low on all others. Then, the underlying dimension is "named" by characterizing the items which load highest on it. Typical uses of factor analysis are exploration of the pattern of relationships which exist between different variables, confirmation of hypothesized relationships which exist between different variables, and the construction of scales which have greater reliability and generality than the individual items. In this study the factor pattern underlying students' evaluations had already been determined (Marsh, 1976a), and a set of 24 evaluation items designed to measure six evaluation factors had been developed. However, faculty self-evaluations had not been previously factor analyzed. The purpose of this analysis was to determine if the factors underlying the students' evaluations were replicable and if they were similar to those underlying faculty self-evaluations of their own teaching.

The set of 24 evaluation items and the factors which they were designed to define are presented in Table One. The factor loadings of the students' evaluations and the faculty self-evaluations both offered support for these six evaluation factors; items loaded higher on the factors they were designed to measure than on other factors. There was only disagreement on two items; student responses to the item "answers questions carefully" put the item in the ORGANIZATION factor, while the faculty responses placed it in the LEARNING/VALUE factor; student responses to the item "discusses recent developments" placed the item in the BREADTH factor, while faculty responses placed it in the ORGANIZATION factor.

- - - - - - - - - - - - - - - - -

INSERT TABLE ONE ABOUT HERE

- - - - - - - - - - - - - - - - -

In summary, factor analyses of both the students' evaluations and the faculty self-evaluations supported the existence of the same six evaluation factors the items were designed to measure. The similarity in the factor patterns implies that the two groups agree upon the dimensions which underlie evaluations of effective teaching. This analysis does not show that the actual ratings which students give a teacher will agree with those which the teacher gives himself. In the next section the student evaluation factors, scores representing the six factors based upon the student ratings, are correlated with the corresponding faculty self-evaluation factors.

Convergent-Divergent Validity

Campbell and Fiske (1959) advocate the assessment of validity by determining measures of more than one trait, each of which is assessed by more than one method. In the present application, the multiple traits are the six evaluation factors while the multiple methods refer to the two distinct groups of raters--the students and the faculty. Convergent validity, that which is most typically determined, is the correlation between the same evaluation factor rated by the two different groups. Discriminant validity refers to the distinctiveness of each of the evaluation factors.

Convergent and divergent validity were determined by examining the set of correlation matrices in Table Two. The two triangular matrices contain the correlations between different evaluation factors as assessed by the same group of raters; intercorrelations between student evaluation factors (upper left) and faculty -evaluation factors (lower right). The diagonals of these triangular matrices contain the reliabilities of the factors for each group of raters. The square matrix (lower left) contains the correlations between student evaluation factors and fac ty self-evaluation factors. The diagonal of the square matrix, the convergent validity coefficients, are the correlations between the same evaluation factors assessed by the two groups of raters. Since there was substantial unreliability in many of the faculty self-evaluation factors, the convergent validity coefficients have been corrected for unreliability (Nunnally, 1967).

Convergent validity requires that the diagonal values of the square matrix be substantially higher than zero. Inspection of Table Two shows that this was the case for all evaluation factors; there was substantial greement

between students and faculty. These validity coefficients, corrected for unreliability, varied between .33 and .67 (median r = .49). These findings offered good support for convergent validity of students' evaluations.

- - - - - - - - - - - - - - -

INSERT TABLE TWO ABOUT HERE

- - - - - - - - - - - - - - -

Divergent validity was much harder to assess, and Campbell and Fiske (1959) offered only general guidelines. The minimal condition is that all correlations between different factors rated by the same group (off-diagonal correlations in the triangular matrices) must be substantially lower than the reliabilities of these factors. For example, even though the correlation between student ratings of ENTHUSIASM and GROUP INTERACTION was .54, this correlation was still much lower then either of the reliabilities of these two factors (.92 and .93 respectively). This first condition was clearly met for both student and faculty ratings. A second condition is that each convergent validity coefficient must be higher than any other correlation in the same row or column of the square matrix. This condition was also met in all cases. A third condition is that a similar pattern of correlations exist in each of the triangular matrices and the square matrix. This was generally the case, particularly if only the correlations which were statistically significant are considered. A final condition, the most stringent, suggests that each convergent validity coefficient should be higher than correlations between that factor and any other factor assessed by the same group of raters. This condition implictly assumes that the evaluation factors are truly uncorrelated, clearly not the case for teaching evaluation factors, and so is only somewhat relevant. This condition was met for the faculty self-evaluations, but was only partially met by the students' evaluations. This suggests that there may be some "halo effect" in the students' evaluations.

In summary, there was very good support for the convergent validity of the teacher evaluations, and reasonably good support for their divergent validity. Student-faculty agreement on the same evaluation factors was quite high; validity coefficients corrected for unreliability varied between .33 and .67 (median r = .49). Furthermore, the agreement was specific to student and faculty ratings on the same evaluation factor. For example, students' evaluations of ORGANIZATION correlated highly with faculty self-evaluations of ORGANIZATION, but did not correlate substantially with any other factors. Reliabilities of the student evaluation factors were high (median r = .90), but were lower for faculty self-evaluation factors (median r = .70). Correlations between different student evaluation factors were somewhat higher than is desirable (median r = .39), perhaps indicating some halo effect, but were substantially less than the reliabilities of the factors.

Student-Faculty Agreement--Absolute and Relative

Results in the last section indicated that the correlations between student evaluation factors and faculty self-evaluation factors were quite high. However, this does not imply there was absolute agreement since correlations can only assess relative agreement. For example, if each instructor always rated himself exactly one category higher than did his students, there would be perfect relative agreement (a correlation of 1.0) even though there would not be absolute agreement. The purpose of analysis in this section is to test both relative and absolute agreement of individual evaluation items.

Mean faculty self-evaluations were very similar to the mean of students' evaluations (See Table Three). For the 24 evaluation items, the median

rating was exactly the same for both groups; 4.07 on the five-point response scale. Differences between faculty and student ratings only reached statistical significance on five items; students' evaluations were higher on two and lower on three.

The mean faculty self-evaluation on each of the 24 evaluation items, averaged across all faculty responses, correlated quite highly ($r = .77$) with the mean student ratings. This implies that students and faculty agree upon what the faculty as a whole does best and worst. For example, both faculty and students rate faculty as most effective at being enthusiastic, being friendly to students, enjoying teaching, being well prepared, and having an interest in students, but perceive faculty to be least effective at giving lectures which are easy to outline, knowing when students are bored and confused, and enhancing presentations with the effective use of humor.

Correlations between faculty self-evaluations and ratings by their students are presented in Table Three. These correlations are similar in meaning to those presented in Table Two, but show agreement on individual items rather than factors. Correlations were significantly positive on 23 of 24 evaluation items, the median correlation being $r = .30$. It is interesting to note that this is lower than the median correlation for factor scores, even before they were corrected for unreliability (median uncorrected $r = .39$). The higher correspondence between the evaluation factors was primarily due to the greater reliability of the factors, compared to the reliability of individual items.

In summary, these findings indicate that there was good agreement--both absolute and relative agreement--between students' evaluations and the corresponding evaluations by their teachers. Differences between mean student

ratings and mean faculty self-evaluations were small, correlations between ratings by the two groups were statistically significant on 23 of 24 items, and there was good agreement between the two groups about what teaching behaviors were more descriptive and less descriptive of the faculty as a whole.

Differentiation Between "Most" and "Less" Effective Courses

Faculty in this study, unlike others which were discussed, selected one course in which their teaching was "most effective" and another course in which their teaching was "less effective". Many potential problems inherent in the use of category ratings were avoided with this procedure. While category ratings are also the basis of students' evaluations, each student is exposed to a variety of different teachers, and the evaluation of each teacher is based upon the responses of many different students. On the other hand, faculty self-evaluations are based upon the response of a single individual who may not have taken an undergraduate course for a decade or more. The self-rating of "4" by one instructor may or may not be different in meaning from the "3" by another instructor. However, if the same instructor gives himself a "4" in one class and a "3" in another, it is clear that he feels that there is a difference between the two classes. This methodology also forces faculty to evaluate critically the difference in their teaching effectiveness in the two courses.

This procedure does present a much more demanding test of the students' evaluations. While the difference between a very good teacher and very poor one may be readily apparent, the difference in teaching effectiveness of the same instructor in two different classes is much more subtle. Furthermore, virtually all of the faculty taught only two courses during the semester, so there was a limited range from which to select a "most effective" and "less effective" course. In fact, a number of faculty (22%) indicated there was

little or no difference in their effectiveness in the two courses even though they did indicate one as "most effective."

Students' evaluations significantly differentiated between the "most effective" and "less effective" courses (See Table Four). Differences on each of the eight evaluation scores separately and the multivariate difference based upon the entire set of evaluation scores were all statistically significant. The largest differences were for the Overall Instructor rating and the Instructor Enthusiasm factor. Faculty self-evaluations also differentiated between the two groups of courses, but differences were not statistically significant for all of the evaluation scores. The "most effective" and "less effective" courses were also compared to 10 background/demographic variables which describe the instructor, the student, and the course. Differences between the two groups of courses failed to reach statistical significance on nine of the ten variables (See Table Five), as well as the multivariate difference based upon all ten variables.

- - - - - - - - - - - - - - - - - - - -

INSERT TABLES FOUR AND FIVE ABOUT HERE

- - - - - - - - - - - - - - - - - - - -

In summary, faculty who had taught two courses during the same semester were asked to indicate the course in which their teaching had been "most effective" and "less effective". Students' evaluations of the "most effective" courses were significantly higher for all evaluation scores, even though the two groups of courses were similar on 10 background/demographic variables. Inspection of Table Four indicated that the students' evaluations actually show better differentiation than did faculty self-evaluations of their own teaching.

DISCUSSION

Faculty who taught two undergraduate courses during the same semester evaluated their own teaching in each course as well as being evaluated by their students. Both faculty and students used essentially the same evaluation forms. In spite of Faculty skepticism concerning the validity of students' evaluation, there was very good student/faculty agreement. Separate factor analyses of the two sets of ratings indicated student-faculty agreement on the evaluation dimensions which underlie the ratings of teaching effectiveness. Validity coefficients, correlations between student ratings and faculty ratings on the same evaluation factors, were all highly significant (median r = .49). The reliabilities of the students' evaluations were high (median r = .90) though faculty self-evaluations were somewhat less reliable (median r = .70). Mean faculty self-evaluations, averaged across all faculty responses, were generally similar to the mean students' evaluations; the median rating for the 24 evaluation items was 4.07 for both groups. Furthermore, there was faculty-student agreement upon which teaching behaviors were more descriptive and less descriptive of the faculty as a whole (r = .77). Finally, when faculty indicated that their teaching was relatively more effective in one of the two courses in which they evaluated themselves, students' evaluations of the "most effective" courses were significantly higher for each of the evaluation factors and both the overall summary ratings. In fact, students' evaluations better differentiated between courses in which faculty indicated that their teaching was "most effective" and "less effective" than did the faculty self-evaluations of their own teaching.

Previous research (Centra, 1972; Blackburn & Clark, 1975) reported lower validity coefficients than were found in this study and also reported that faculty self-evaluations were consistently higher than were students' evaluations. The differences in their findings may well be due to the lower reliability of the measurement instruments which they used and the different methodologies which they employed. Blackburn and Clark (1975), reporting the lowest level of agreement, had each teacher rate himself on a single global item of "overall teaching effectiveness" which was not actually tied to performance in a particular course. Reliability data was not presented for either the students' evaluations, also assessed with a single item, or the faculty self-evaluations. Centra (1972) asked faculty to select one course in which they would evaluate themselves and be evaluated by their students at the middle of the semester. The effect of the timing of the evaluations, coming at the middle of the semester, is not known, but this is an obvious difference in methodology. Furthermore, many faculty in the study taught more than one course and probably selected the course in which they felt their teaching was most effective. If this selection bias did exist, it would produce more favorable faculty self-evaluations, it would limit the range of teaching effectiveness which was actually observed (lowering the correlations between the two sets of evaluations), and it would not force faculty to differentiate between their own more effective and less effective teaching. Although Centra did include items which apparently tapped different evaluation factors, validity coefficients were based upon individual evaluation items rather than evaluation factors. Reliabilities of Centra's items (Centra, 1973) were somewhat lower than reliabilities of individual items in the present study and were markedly lower than the reliabilities of the evaluation factors. Centra presented no data on the reliabilities of the faculty self-evaluations.

Faculty in this study, as is generally the case, were somewhat skeptical about the validity of students' evaluations and their use. Faculty did indicate that some measure of teaching effectiveness should be given more weight in tenure/promotion decisions and did indicate that the students' evaluations were useful to the faculty themselves. (78% agreed with the statement that "students' evaluations can provide instructors with information which is useful for the improvement of the course and/or quality of teaching", while only 2% disagreed.) However, Faculty were skeptical about the validity of students' evaluations. (Only 28% agreed with the statement that "students' evaluations represent accurate assessments of instructional quality".) Similar reservations were expressed about the use of students' evaluation in tenure/promotion decisions. Furthermore, faculty were equally skeptical about other possible measures of teaching effectiveness which were suggested and did not provide any alternative measures which met with their approval. A dilemma clearly exists. Faculty are concerned about teaching effectiveness, even to the extent of wanting it to play a major role in administrative decisions, but have no confidence in any measures of teaching effectiveness--including students' evaluations. Before the potential usefulness of students' evaluations can be achieved, faculty and administrators (who are generally faculty, former faculty, or least "faculty-like" people) have to be willing to trust students' evaluations.

An important role of research in students' evaluations, besides demonstrating their reliability, validity and lack of bias, is to convince faculty and administrators of their worth. No matter how good a measure actually is, it is of little value unless it is used. Previous research has clearly demonstrated the validity of students' evaluations against a multitude of criteria, yet faculty are still skeptical. Any particular validity criterion is generally either quite specific to a particular course (e.g., standardized performance in multi-section course of calculus or computer programming) or can be attacked as being inappropriate (e.g., alumni ratings). In the present study, the criterion used was faculty self-evaluations of their own teaching.

In spite of the faculty reservations about the accuracy of the students' evaluations, the results showed good student-faculty agreement. Not only does this study provide important new evidence for the validity of students' evaluations, but the findings should be instrumental in overcoming faculty reservations about the students' evaluations.

Faculty self-evaluations in this study have been used primarily as a criterion for validating students' evaluations. However, the findings do suggest that under some circumstances the faculty self-evaluations can be useful as well. Factor analysis of the faculty self-evaluations gave evidence of a well defined factor structure. The faculty self-evaluations gave showed good agreement with students' evaluations. Even the moderate lack of reliability of the faculty self-evaluations might be overcome if data were averaged across several different courses. While it is probably unrealistic to expect faculty to be objective if their self-evaluations were to be used for tenure/promotional decisions, their ratings may prove quite valuable to the improvement of teaching. The thought processes necessary to complete the self-evaluations require that faculty carefully scrutinize their teaching. Furthermore, Centra (1972) reported that faculty who found that their students evaluated them much lower than they had evaluated themselves were more likely to benefit from the feedback provided by students' evaluations.

REFERENCES

Blackburn, R.T. and Clark, M.J. An assessment of faculty performance: Some correlates between administrators, colleagues, student, and self ratings. Sociology of Education, 1975, 48, 242-256.

Campbell, D.T., and Fiske, D.W. Convergent and discriminant validation by the multi-trait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Centra, J.A. Self-Ratings of College Teachers: A Comparison with Student Ratings. (Student Instructional Report No. 2) Princeton, N.J. Educational Testing Service, 1972.

Centra, J.A. The Student Instructional Report: Item Reliabilities (Student Instructional Report No. 3) Princeton, Service, 1973.

Centra, J.A. Colleagues as raters of classroom instruction. Journal of Higher Education, 1975, 46, 327-337.

Centra, J.A. Student ratings of instruction and their relationship to student learning. American Educational Research Journal, 1977, 14, 17-24.

Cohen, S.A. and Berger, W.G. Dimensions of students' ratings of college instructors underlying subsequent achievement on course examinations. Proceedings of the 78th Annual Convention of the American Psychological Association, 1970, 5, 605-606. (Summary)

Frey, P.W. Student ratings of teaching: Validity of several rating factors. Science, 1973, 182, 83-85.

Guthrie, E.R. The Evaluation of Teaching: A Progress Report. Seattle: University of Washington, 1954.

Hildebrand, M., Wilson, R.C., and Dienst, E.R. Evaluating University Teaching. Berkeley: Center for Research and Development in Higher Education, University of California, Berkeley, 1971.

Marsh, H.W. Reliabilities of Items on the Student Evaluation Form Used in the Social Science Division at USC. Los Angeles: Office of Institutional Studies, University of Southern California, 1976a. (OIS 76-8).

Marsh, H.W. Factor Analysis of the Student Evaluation Form Used in the Social Science Division at USC. Los Angeles: Office of Institutional Studies, University of Southern California, 1976b. (OIS 76-7).

Marsh, H.W. The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. American Educational Research Journal, 1977, 14, 441-447.

Marsh, H.W. Students' Evaluations of Instructional Effectiveness: Relationship to Student, Course, and Instructor Characteristics. Paper presented at Annual Meeting of American Educational Research Association, Toronto, March, 1978.

Marsh, H.W., Fleiner, H., and Thomas, C.S. Validity and usefulness of student evaluations of instructional quality. Journal of Educational Psychology, 1975, 67, 833-839.

Marsh, H.W., Overall, J.U., and Thomas, C.S. The Relationship Between Students' Evaluation of Instruction and Expected Grade. Paper presented at Annual Meeting of American Educational Research Association, San Francisco, April, 1976. (ERIC ED 126 140).

Maslow, A.H., and Zimmerman, W. College teaching ability, scholarly activity, and personality. Journal of Educational Psychology, 1956, 46, 185-189.

McKeachie, W.J. Correlates of Student Ratings. Proceedings: The First Invitational Conference on Faculty Effectiveness as Evaluated by Students, A.L. Sockloff, ed. Philadelphia: Measurement and Research Center, Temple University, 1973.

Morsh, J.E., Burgess G.G. and Smith, P.N. Student achievement as a measure of instructor effectiveness. Journal of Educational Psychology, 1956, 47, 79-88.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., and Bent, D.H. Update to Statistical Package for the Social Sciences. New York: McGraw-Hill Book Company, 1977.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K. and Bent, D.H. Statistical Package for the Social Sciences. New York: McGraw-Hill Book Company, 1975.

Nunnally, J.C. Psychometric Theory. New York: McGraw-Hill, 1967.

Overall, J.U. and Marsh, H.W. Cognitive and Affective Outcomes: Their Relationship to Effective Teaching and Students' Evaluations. Paper presented at Annual Meeting of American Educational Research Association, Toronto, March, 1978.

Remmers, H.H. Teaching methods in research on teaching. Handbook of Research on Teaching, N.L. Gage, ed. Chicago: Rand McNally, 1963.

Rodin, M. and Rodin, B. Student evaluations of teachers. Science, 1972, 177, 1164-1166.

TABLE ONE
Factor Analyses of Students' Evaluations and Faculty Self-Evaluations
(N=83 Courses)

	I		II		III		IV		V		VI	
I. BREADTH OF COVERAGE												
Discusses other points of view	**71**	**(59)**	10	(13)	27	(19)	13	(-11)	04	(-21)	-09	(-05)
Contrasts implications	**86**	**(86)**	14	(06)	-02	(-15)	-07	(-09)	0	(23)	06	(07)
Discusses recent developments	**44**	**(23)**	-23	(44)	10	(12)	10	(16)	29	(-14)	24	(24)
Presents origins of ideas/concepts	**51**	**(62)**	10	(-08)	13	(-05)	22	(32)	07	(17)	26	(18)
ORGANIZATION												
Explains clearly	06	(-10)	**73**	**(77)**	02	(-12)	01	(-09)	33	(15)	09	(12)
Is well prepared	25	(17)	**80**	**(51)**	-17	(-01)	09	(16)	-17	(23)	18	(11)
Lectures easily outlined	-03	(21)	**70**	**(73)**	14	(-12)	10	(-04)	14	(-03)	08	(04)
Answers questions carefully	17	(11)	**60**	**(04)**	30	(18)	15	(-09)	08	(-16)	03	(75)
III. GROUP INTERACTIONS												
Encourages class discussions	02	(03)	03	(-23)	**88**	**(85)**	-03	(09)	02	(11)	12	(-02)
Invites sharing of knowledge/ideas	07	(06)	-07	(-04)	**86**	**(84)**	15	(-02)	09	(-06)	-02	(01)
Invites criticism of own ideas	22	(51)	09	(02)	**65**	**(51)**	29	(06)	06	(-14)	01	(-00)
Knows when students confused/bored	-02	(-03)	22	(12)	**40**	**(48)**	19	(-18)	32	(22)	21	(21)
IV. INDIVIDUAL RAPPORT												
Has interest in students	-02	(00)	12	(30)	28	(06)	**69**	**(44)**	20	(12)	-01	(11)
Is friendly to students	08	(-18)	05	(24)	25	(35)	**78**	**(35)**	23	(06)	-20	(-02)
Is accessible out of class	18	(02)	-03	(-07)	-04	(00)	**67**	**(41)**	-21	(-06)	22	(14)
V. INSTRUCTOR ENTHUSIASM												
Is dynamic and energetic	14	(-04)	18	(08)	-02	(02)	23	(10)	**48**	**(80)**	30	(07)
Has interesting presentation style	05	(02)	25	(28)	08	(11)	-03	(-12)	**63**	**(57)**	27	(24)
Is enthusiastic about subject	10	(04)	42	(35)	-17	(04)	36	(40)	**30**	**(65)**	08	(-14)
Inhances presentation with humor	16	(07)	24	(17)	-05	(25)	24	(51)	**38**	**(49)**	30	(-02)
OVERALL COURSE RATING	16	(-04)	-07	(32)	12	(29)	06	(-38)	**74**	**(29)**	-05	(16)
	11	(12)	27	(02)	15	(05)	23	(-23)	**38**	**(64)**	30	(19)
VI. LEARNING												
Course intellectually demanding	11	(-04)	03	(-12)	00	(-18)	12	(23)	-09	(14)	**72**	**(61)**
You learned something valuable	06	(02)	06	(18)	13	(-15)	-04	(27)	20	(00)	**68**	**(54)**
OVERALL COURSE RATING	12	(22)	19	(02)	19	(05)	-07	(-16)	22	(44)	**61**	**(44)**

1-Factor loadings in bold boxes are loadings for items designed to measure each Factor. Results of factor analysis Faculty self-evaluations are presented in parenthises.

2-Factor Analyses of both sets of data consisted of a principle components analysis, kaiser normalization, and rotation to a direct oblimin criterion for which the delta parameter was set at--2.0. Analysis was performed with the commerically available statistical Package for Social Scientists (Nie. et. al, 1975).

Reliabilities and inter correlations between the different factors are presented in the next section.

TABLE TWO

CONVERGENT AND DISCRIMINANT VALIDITY
(N=83 courses evaluated by both students and faculty)

STUDENTS' EVALUATIONS	Student's Evaluations (1)	(2)	(3)	(4)	(5)	(6)	Faculty Self-Evaluations (7)	(8)	(9)	(10)	(11)	(12)
(1) ENTHUSIASM	(92)											
(2) GROUP INTERACTION	54	(93)										
(3) LEARNING	24	04	(87)									
(4) INDIVIDUAL RAPPORT	39	46	05	(86)								
(5) BREADTH OF COVERAGE	46	35	45	44	(86)							
(6) ORGANIZATION	61	28	35	39	47	(93)						
FACULTY SELF-EVALUATIONS												
(7) ENTHUSIASM	37(.42)	16	15	02	12	17	(85)					
(8) GROUP INTERACTION	00	47(.55)	-17	-01	00	-12	22	(79)				
(9) LEARNING	07	-18	38(.50)	04	16	05	21	-17	(67)			
(10) INDIVIDUAL RAPPORT	-06	-09	29	32(.48)	07	10	09	-04	21	(51)		
(11) BREADTH OF COVERAGE	-03	-16	20	-03	26(.33)	05	09	18	29	04	(72)	
(12) ORGANIZATION	28	-13	36	07	22	55(.67)		-19	14	24	21	(72)

1-Values in diagonals of upper left and lower right matricies, the two triangular matricies, are reliability estimates (coefficient alphas). (See Nie, et. al., 1977)

2-Values in diagonals of lower left matrix, the square matrix, are validity coefficients. The values in parentheses have corrected for unreliability according to the equation: corrected r_{xy} = (uncorrected r_{xy})/$\sqrt{(r_{xx})(r_{yy})}$

3-Correlation coefficients are presented without decimal points; correlations greater than 20 are statistically significant.

TABLE THREE

Agreement Between Faculty Self-Evaluations and Evaluations of Their Students
(N=83 courses evaluated by both faculty and students)

	ABSOLUTE AGREEMENT		RELATIVE AGREEMENT
	Mean Faculty Self-Evaluations	Mean Student Evaluations	Correlation Between Faculty and Student Evaluations
I. BREADTH OF COVERAGE			
Discusses other points of view	4.04	4.12 NS	+ .19 *
Contrasts implications	3.96	4.19 *	+ .32 **
Discusses recent developments	4.30	4.16 NS	+ .27 **
Presents origins of ideas/concepts	4.04	4.11 NS	+ .21 *
II. ORGANIZATION			
Explains clearly	3.95	3.99 NS	+ .49 **
Is well prepared	4.36	4.24 NS	+ .42 **
Lectures easily outlined	3.58	3.69 NS	+ .44 **
Answers questions carefully	4.18	4.01 NS	+ .05 NS
III. GROUP INTERACTION			
Encourages class discussions	4.11	4.07 NS	+ .47 **
Invites sharing of knowledge/ideas	3.76	3.95 NS	+ .47 **
Invites criticism of own ideas	3.94	3.72 NS	+ .23 *
Knows when students confused	3.80	3.54 *	+ .19 *
IV. INDIVIDUAL RAPPORT			
Has interest in students	4.74	4.17 **	+ .20 *
Is friendly to students	4.52	4.39 NS	+ .21 *
Is accessible out of class	4.06	4.11 NS	+ .50 **
V. INSTRUCTOR ENTHUSIASM			
Is dynamic and energetic	4.17	3.95 *	+ .28 **
Has interesting presentation style	3.82	3.78 NS	+ .35 **
Enjoys teaching	4.26	4.49 **	+ .46 **
Is enthusiastic about subject	4.60	4.45 NS	+ .26 **
Enhances presentations with humor	3.64	3.81 NS	+ .41 **
OVERALL COURSE RATING	4.08	4.07 NS	+ .31 **
VI. LEARNING			
Course intellectually demanding	4.23	4.07 NS	+ .23 *
You learned something valuable	4.11	4.23 NS	+ .29 *
OVERALL COURSE RATING	3.79	3.82 NS	+ .32 *
(Median of 24 items)	(4.07	4.07)	(+ .30)

* p .05 **p .01 NS-Not Signficant

1- Two-tailed statistical tests were used in determining absolute agreement since it was assumed that students' evaluations may be either higher or lower than faculty self-evaluations. One-tailed tests were used to test relative agreement since it was assumed that correlations would only be positive.

2- The correlation between the 24 mean faculty responses and the 24 mean student responses is .77 indicating good agreement on what teaching behaviors are more or less descriptive of faculty as a whole.

3- The correlations between faculty and student responses were not corrected for unreliability which is substantial for faculty self-evaluations.

4- Evaluation factor scores were not used in this analysis since factors scores, a weighted average of z-scores, have a mean of 0.0 (or some other arbitrary value).

TABLE FOUR

DIFFERENCES IN EVALUATIONS OF COURSES IN FACULTY INDICATED THEIR TEACHING WAS "MOST EFFECTIVE" AND "LESS EFFECTIVE" (N=32 "most effective' and 32 "less effective" courses)

Evaluation Factors[1]	Students' Evaluations		Faculty Self-Evaluations	
	Most Effective Courses	Less Effective Courses	Most Effective Courses	Less Effective Courses
ENTHUSIASM	105.3	96.9 **	102.1	98.9 *
GROUP INTERACTION	104.5	99.0 **	102.3	99.0 NS
INDIVIDUAL RAPPORT	103.3	98.7 **	100.7	102.1 NS
BREADTH OF COVERAGE	103.7	66.2 **	100.2	97.2 NS
VALUE/LEARNING	102.6	98.2 *	104.7	97.2 **
ORGANIZATION	103.2	98.9 *	101.6	97.9 *
OVERALL INSTRUCTOR	4.24	3.98 **	4.27	3.97 **
OVERALL COURSE	3.96	3.74 **	4.11	3.57 **

* $p < .05$; ** $p < .01$, NS--Not Significant

1- Evaluation factors, the first six evaluation scores, were standardized (mean=100, standard deviation=15) for students and faculty separately. The two Overall Summary items varied along a five-point scale ranging from "1-Among the Worst" to "5-Among the Best"

2- Statistical significance was determined by a one-tailed dependent t-test, since scores were predicted to be higher in the "most effective" courses on a prior basis.

3- Multivariate significance tests, Hotelings T-Squared, indicated significant differentiation between the two groups of courses with both the students' evaluations (Hotelling T-Square = 33.2 ; $F(8,24) = 3.3$, $p < .01$) and Faculty self evaluations (Hotelling T-Square = 31.6 ; $F(8,22) = 2.9$ $p < .05$).

4- Results presented in this table based upon only Faculty who rated themselves and were rated by their students in two undergraduate courses. A total of 64 courses, 32 pairs, met this criteria. The remaining 19 courses were either paired with a graduate level course, or were unpaired (i.e. Faculty rated only one course).

TABLE FIVE

BACKGROUND/DEMOGRAPHIC DIFFERENCES BETWEEN "MOST EFFECTIVE" AND "LESS EFFECTIVE" COURSES

(N=32 "most effective" & "less effective")

Background/Demographic Variable	Most Effective Courses	Less Effective Courses	
Number of Times Instructor Had Taught Same or Similar Course	5.24	4.17	NS
Faculty Impressions of Student Interest in Subject at Start of the Course (1-Very Low...5-Very High)	3.41	3.17	NS
Instructor's Self-Rating of "Grading Leniency" (1-Very Easy Grader...5-Very Hard/Strict Grader)	3.57	3.57	NS
Class Average of Students' Expected Grade (0-F....4-A)	3.33	3.27	NS
Average of Students' GPA (1-Below 2.4, 2-2.4 2-2.4 to 2.9 3-2.9 to 3.37, 4-3.37 to 3.7, 5-Above 3.7)	3.46	3.42	NS
Percentage of Upper Division Students in Class	60%	50%	NS
Percentage of Students Majoring in Division	59%	48%	NS
Percentage of Studetns Taking Course to "Fulfill a Major Requirment"	51%	46%	NS
Course Level (1-Lower Division, 2-Upper Division)	1.84	1.66	*

** p .01; * p .05; NS-Not Significant

1-Statistical significance was tested with two-tailed dependent t-tests since there was no apriori basis for predicting the direction of the differences

2-Multivariate significance, Hotelings T-Square, indicated that assessing all 10 variables simultaneously, the differences between the two groups of courses was not statistically significant (Hotelling T-Square= 17.6; $F(10,20)= 1.21$ $p > .05$)